Машинное обучение для решения исследовательских и инженерных задач в науках о Земле

Михаил Криницкий

к.т.н., н.с.
Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)

Классификация задач и методов машинного обучения

Михаил Криницкий

к.т.н., н.с.
Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

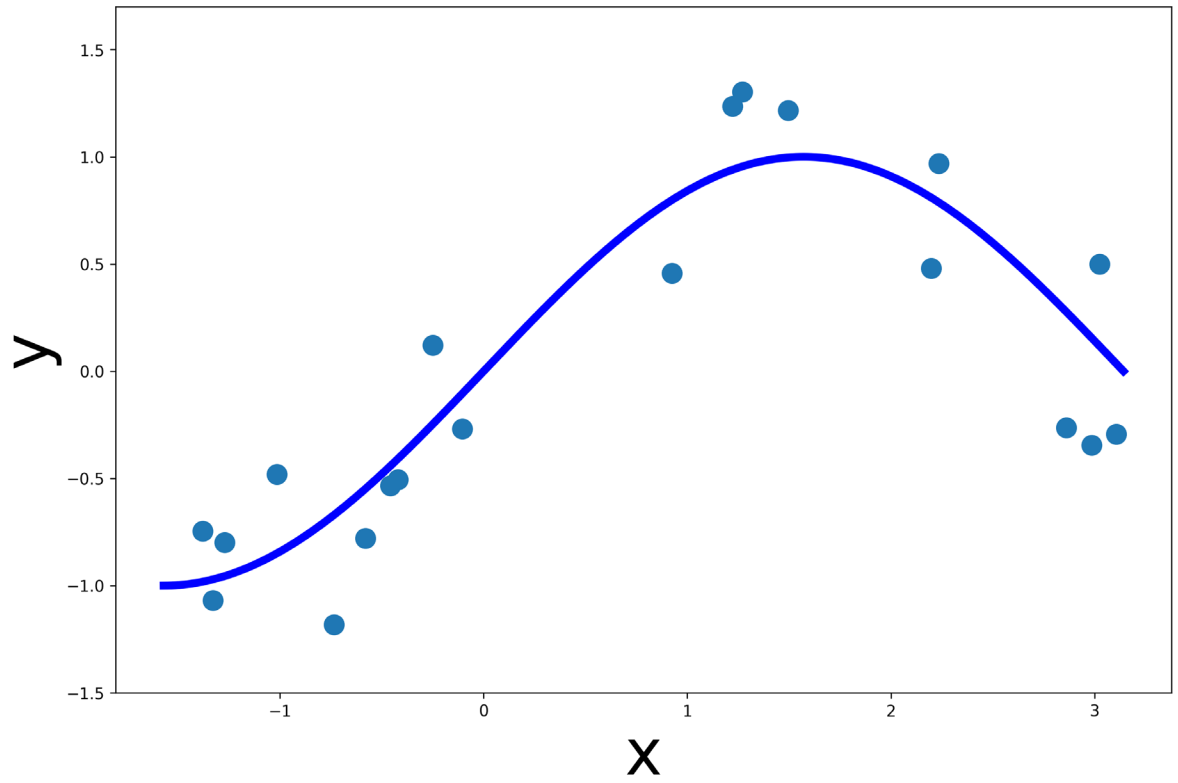
ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

- «Обучение с учителем»
  - восстановление регрессии

что я хочу? – значение y

$y \in \mathbb{R}^m$

m – размерность целевой переменной



ремарки

- Количество размеченных данных (зачастую) играет роль
- Разные модели ведут себя по-разному в зависимости от шума в данных, от количества данных, от наличия выбросов в данных
- Сложная точная модель – не обязательно лучшая для конкретной задачи

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

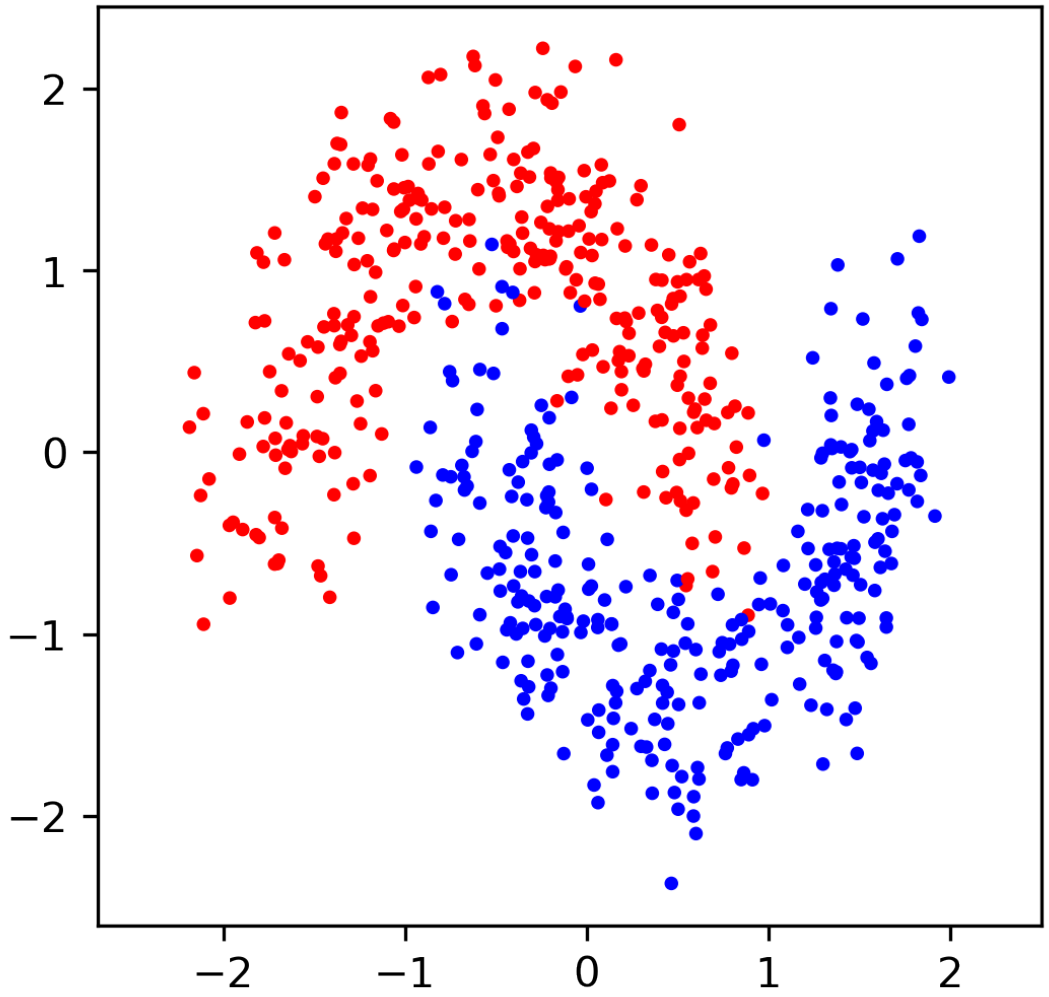
ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

- «Обучение с учителем»
  - восстановление регрессии
  - классификация

что я хочу? – метку класса

«красный или синий?»

(бинарная классификация)



КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

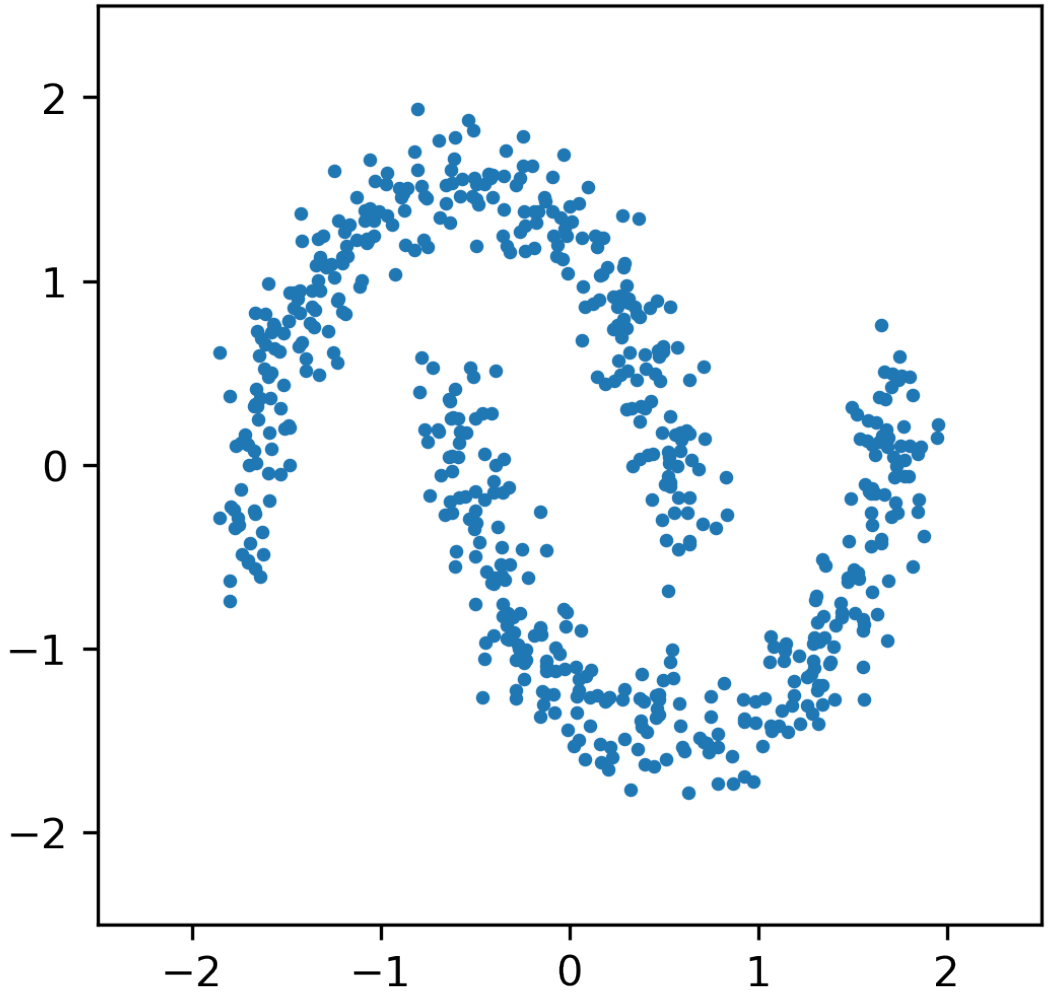
ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

типы задач:

- «Обучение с учителем»
  - восстановление регрессии
  - классификация
- «Обучение без учителя»
  - поиск структуры в данных

что я хочу?

- метки групп
- знать, есть ли группы?
- сколько групп?



ремарки

Кластеризация:

- Количество данных часто, но не всегда играет роль

- Разные модели ведут себя по-разному в зависимости от шума в данных, от количества данных, от наличия выбросов в данных, от наличия структуры в данных

- Разные модели дают разный результат, но нет «более правильного» результата. Есть «более подходящий» для целей конкретного исследования.

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

типы задач:

- «Обучение с учителем»
  - восстановление регрессии
  - классификация
- «Обучение без учителя»
  - кластеризация
  - понижение размерности

что я хочу?

- новое представление (признаковое описание) данных в пространстве ме́ньшей размерности

цели:

- Визуализация на плоскости, в 3D
- Борьба с переобучением (в контексте т.н. «проклятия размерности»)
- Сжатие данных с минимальными потерями
- Сокращение вычислительных затрат при обработке данных
- Извлечение значимых признаков, feature engineering

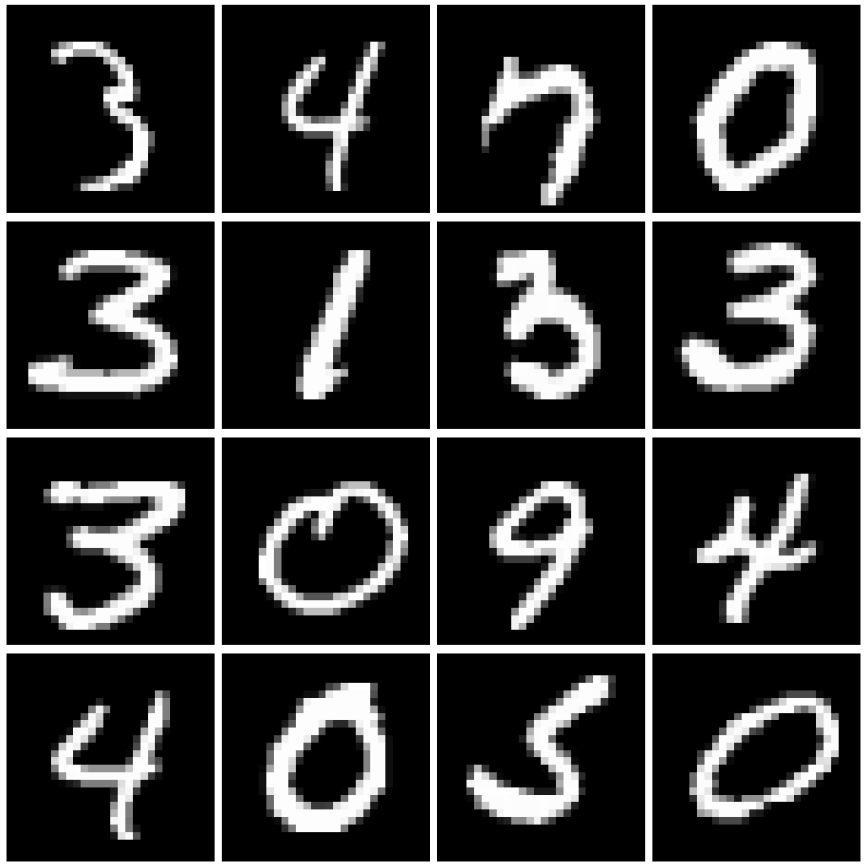
Пожелания:

- Сохранение структуры данных
- Сохранение отношений близости между объектами (событиями)
- Возможность визуализации
- Интерпретируемость новых признаков

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

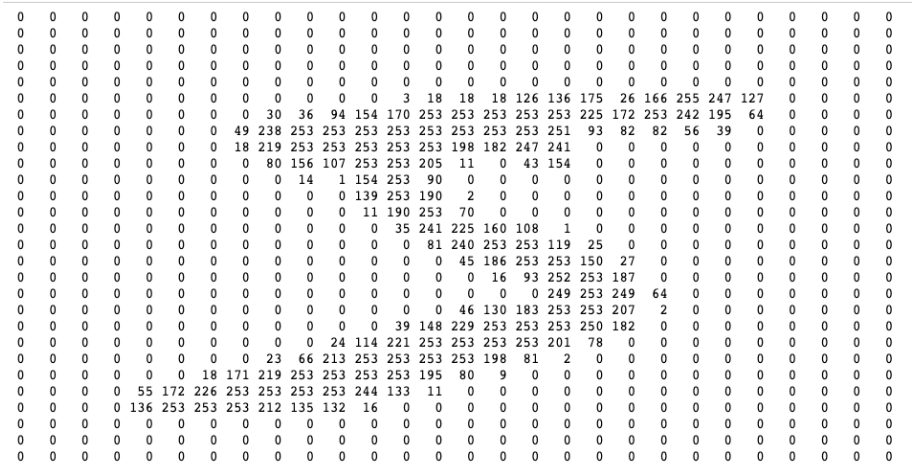
Задача понижения размерности: пример

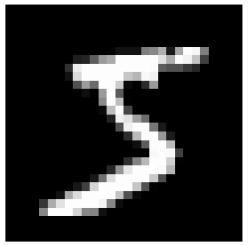
MNIST dataset*



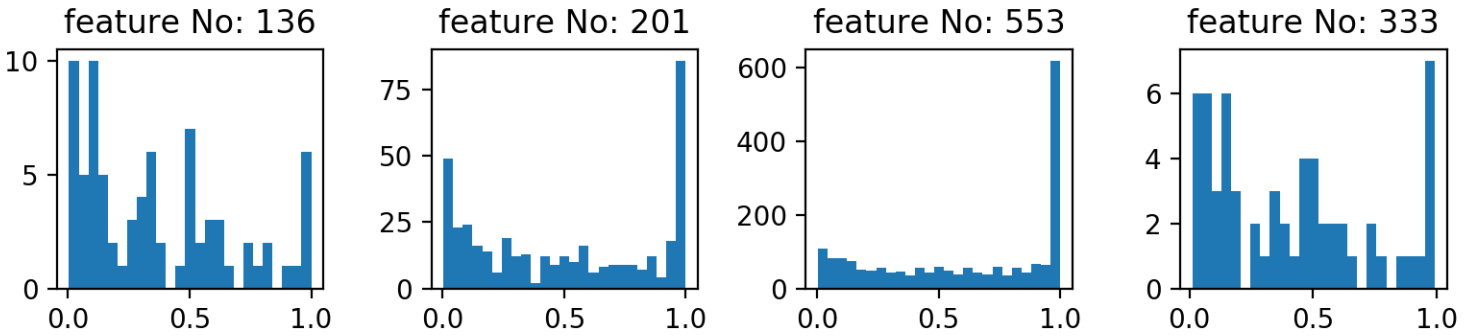
* http://yann.lecun.com/exdb/mnist/

example #0 (label = ”5”)





data distribution (4 of 784 features)



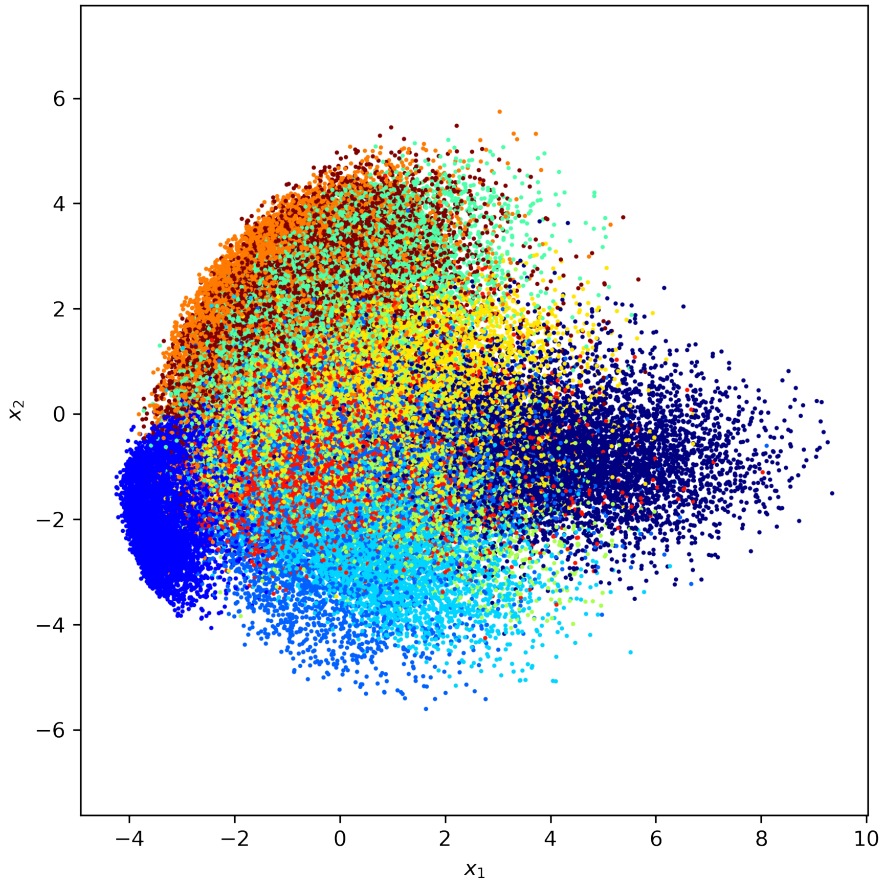
КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

Задача понижения размерности: пример

PCA

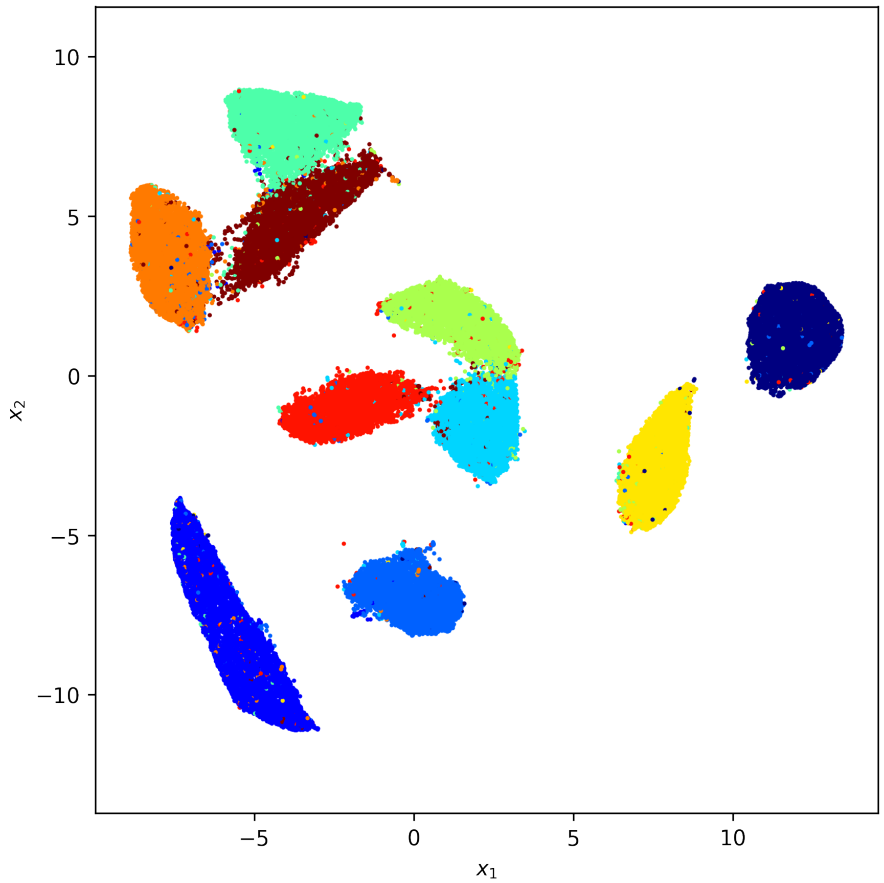
Principal components analysis

Метод главных компонент



UMAP

Uniform Manifold Approximation and Projection



ремарки

Понижение размерности:

- Количество данных почти всегда играет роль

- Разные модели ведут себя по-разному в зависимости от шума в данных, от количества данных, от наличия выбросов в данных, от наличия структуры в данных

- Разные модели дают разный результат, нет «более правильного» результата. Но есть «более подходящий» для целей конкретного исследования.

- Модели различаются по интерпретируемости (e.g. PCA vs. UMAP)

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение с учителем»
  - восстановление регрессии
  - классификация
- «Обучение без учителя»
  - кластеризация
  - понижение размерности
  - восстановление распределения данных

что я хочу?

- Получить модель, генерирующую примеры, распределение которых совпадает с распределением обучающих данных

Цели:

- дополнение данных
- заполнение пропусков в данных

Примеры:

- DeepFake - video
- SuperResolution - images
- Text-to-speech - audio (Siri, Алиса, Cortana, Alexa etc.)

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение с учителем» - **Supervised learning**
  - восстановление регрессии
  - классификация
- «Обучение без учителя» - **Unsupervised learning**
  - кластеризация
  - понижение размерности
  - восстановление распределения данных

другие (реже используемые в ES) типы

- «Обучение с частичным привлечением учителя» - **Weakly supervised learning**
- «Обучение с подкреплением» - **Reinforcement learning**

ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

обучаем (тренируем) модель **на имеющихся данных**

«Обучение»

Объекты

признаки

объекты

	1	2	3	4	5	6	7
1	-0.0685	-0.0676	-0.0643	-0.0617	-0.0550	-0.0488	-0.0365
2	-0.0706	-0.0726	-0.0633	-0.0579	-0.0490	-0.0390	-0.0267
3	-0.0691	-0.0766	-0.0676	-0.0589	-0.0484	-0.0350	-0.0172
4	-0.0631	-0.0719	-0.0662	-0.0557	-0.0441	-0.0317	-0.0144
5	-0.0500	-0.0634	-0.0603	-0.0502	-0.0384	-0.0263	-0.0115
6	-0.0377	-0.0491	-0.0449	-0.0352	-0.0214	-0.0062	0.0056
7	-0.0375	-0.0319	-0.0235	-0.0133	0.0012	0.0169	0.0282
8	-0.0540	-0.0310	-0.0169	6.9177e-...	0.0181	0.0339	0.0445
9	-0.0725	-0.0550	-0.0158	0.0145	0.0323	0.0501	0.0641
10	-0.0789	-0.0844	-0.0512	-0.0010	0.0387	0.0612	0.0810
11	-0.0803	-0.0965	-0.0789	-0.0543	-0.0036	0.0405	0.0685
12	-0.0995	-0.1248	-0.1134	-0.0950	-0.0699	-0.0281	0.0261
13	-0.1026	-0.1306	-0.1283	-0.1171	-0.0982	-0.0806	-0.0412
14	-0.0977	-0.1284	-0.1301	-0.1188	-0.1054	-0.0932	-0.0677
15	-0.0840	-0.1133	-0.1138	-0.1040	-0.0895	-0.0750	-0.0511
16	-0.0820	-0.1089	-0.1072	-0.0967	-0.0814	-0.0669	-0.0464
17	-0.0655	-0.0685	-0.0617	-0.0595	-0.0549	-0.0499	-0.0396

Целевая переменная

	1
1	-0.0685
2	-0.0706
3	-0.0691
4	-0.0631
5	-0.0500
6	-0.0377
7	-0.0375
8	-0.0540
9	-0.0725
10	-0.0789
11	-0.0803
12	-0.0995
13	-0.1026
14	-0.0977
15	-0.0840
16	-0.0820
17	-0.0655

Настройка параметров модели

НАСТРОЕННАЯ МОДЕЛЬ

Модель - метод, структура, значения параметров

$$\hat{y}_i = f(\vec{p}, x_i)$$

Режим использования

новые объекты

	1	2	3	4	5	6	7	8	9
1	-0.0685	-0.0676	-0.0643	-0.0617	-0.0550	-0.0488	-0.0365	-0.0235	-0.0133
2	-0.0706	-0.0726	-0.0633	-0.0579	-0.0490	-0.0390	-0.0267	-0.0136	-0.0044
3	-0.0691	-0.0766	-0.0676	-0.0589	-0.0484	-0.0350	-0.0172	-0.0038	0.0054
4	-0.0631	-0.0719	-0.0662	-0.0557	-0.0441	-0.0317	-0.0144	-0.0022	0.0084
5	-0.0500	-0.0634	-0.0603	-0.0502	-0.0384	-0.0263	-0.0115	0.0020	0.0152
6	-0.0377	-0.0491	-0.0449	-0.0352	-0.0214	-0.0062	0.0056	0.0145	0.0238
7	-0.0375	-0.0319	-0.0235	-0.0133	0.0012	0.0169	0.0282	0.0355	0.0403

НАСТРОЕННАЯ МОДЕЛЬ

Результат:
значения целевой переменной

ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

обучаем (тренируем) модель **на имеющихся данных**

(вернее, на большей их части)

формулировка задачи:

$x \in \mathbb{X}$ — объекты

$y \in \mathbb{Y}$ — ответы

$\mathcal{F}: \mathbb{X} \to \mathbb{Y}$ — искомая закономерность

$\mathcal{T}: \{x_i; y_i\}$ — «обучающая выборка» (прецеденты)

Найти: $\hat{\mathcal{F}}: \{x_i\} \to \{y_i\}$

ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

обучаем (тренируем) модель **на имеющихся данных**

формулировка задачи:

$x \in \mathbb{X}$ — объекты

$y \in \mathbb{Y}$ — ответы

$\mathcal{F}: \mathbb{X} \to \mathbb{Y}$ — искомая закономерность

$\mathcal{T}: \{x_i; y_i\}$ — «обучающая выборка» (прецеденты)

Найти: $\hat{\mathcal{F}}: \{x_i\} \to \{y_i\}$

один из способов решения:

$\mathcal{L}\left(\hat{\mathcal{F}}(x)\right)$ — функционал ошибки (эмпирического риска, потерь)

$\hat{y}_i = \hat{\mathcal{F}}(x_i) = f(\vec{p}, x_i)$ — функционально задаваемая зависимость. **Предположение исследователя о виде закономерности**. Иногда задается параметрически, $\vec{p}$ — вектор параметров.

$\mathcal{L} = L(\vec{p}, \mathcal{T})$ — функция ошибки

$$\hat{p} = \underset{\mathbb{P}}{\operatorname{argmin}}\left(L(\vec{p}, \mathcal{T})\right)$$

$$\hat{\mathcal{F}} = f(\hat{p}, x)$$

ЛИНЕЙНАЯ РЕГРЕССИЯ